Artificial Intelligence PhD Qualifying Examination

High-Fidelity Mock Examination 02

Choose three of four sections; maximum score: 100 points.

Name: _____ **Student ID:** _____ **[1 point]**

Instructions. This examination consists of four sections, each worth 33 points. Choose **three** sections and indicate your choices clearly at the beginning of your answer sheet. If you answer more than three sections, only the first three sections that you select will be graded. The additional point above is awarded for writing your name and student ID. Give complete reasoning unless a question states otherwise.

A. Machine Learning Theory [33 points]

Part I: Warm-Up Questions [3 points]. Select one answer per item.

(i) [0.5 points] Empirical Rademacher complexity conditions on

(a) the sample.

(b) the population risk only.

(c) the optimizer only.

(d) the test labels.

(ii) [0.5 points] If $\mathcal{F}_1 \subseteq \mathcal{F}_2$, empirical Rademacher complexity is

(a) monotone nondecreasing with the class.

(b) always equal.

(c) an integer.

(d) undefined.

(iii) [0.5 points] Agnostic ERM competes with

(a) the best member of its class.

(b) a zero-error classifier.

(c) only the Bayes rule.

(d) a random sign vector.

(iv) [0.5 points] A monotone disjunction predicts one when

(a) at least one retained variable is one.

(b) every variable is zero.

(c) its Gram matrix is PSD.

(d) a margin is negative.

(v) [0.5 points] For a differentiable strongly convex objective, a small gradient

(a) controls distance to the minimizer.

(b) implies infinite VC dimension.

(c) removes curvature.

(d) is a validation label.

(vi) [0.5 points] A bounded-loss validation bound for r candidates contains

(a) $\log r$.

(b) no dependence on r.

(c) 2^r outside a logarithm necessarily.

(d) only training loss.

Part II: Theoretical Exercises [30 points].

7 points For $S_X = (x_1, \dots, x_m)$ define

$$\mathcal{G}_2(B) = \{x \mapsto \langle w, x\rangle : \|w\|_2 \le B\}.$$

Starting from empirical Rademacher complexity, prove

$$\widehat{\mathfrak{R}}_{S_X}(\mathcal{G}_2(B)) \le \frac{B}{m}\sqrt{\sum_{i=1}^{m}\|x_i\|_2^2}.$$

Deduce $BR/\sqrt{m}$ when $\|x_i\|_2 \le R$.

7 points Let $\mathcal{H}$ be finite and let $\widehat{h}$ minimize empirical 0–1 risk. Using Hoeffding's inequality and a union bound, give an explicit sufficient sample size such that, with probability at least $1-\delta$,

$$L(\widehat{h}) \le \min_{h\in\mathcal{H}} L(h) + \varepsilon.$$

All numerical constants must be shown.

5 points Let $F : \mathbb{R}^d \to \mathbb{R}$ be differentiable and λ-strongly convex, with minimizer $w^\star$. Prove

$$\|w - w^\star\|_2 \le \frac{\|\nabla F(w)\|_2}{\lambda}, \qquad F(w) - F(w^\star) \le \frac{\|\nabla F(w)\|_2^2}{2\lambda}.$$

Thus explain what a gradient stopping rule certifies.

6 points On $X = \{0,1\}^d$, define the monotone-disjunction class

$$\mathcal{D}_d = \left\{ h_S(x) = \bigvee_{j\in S} x_j : S \subseteq [d] \right\},$$

where the empty disjunction is constant zero. Determine $\mathrm{VCdim}(\mathcal{D}_d)$.

5 points Consider the bias-free feedforward ReLU network

$$f(x) = W_L\sigma(W_{L-1}\sigma(\cdots\sigma(W_1x))),$$

where $\sigma(u) = \max\{u, 0\}$ coordinatewise. Prove

$$|f(x) - f(x')| \le \left(\prod_{\ell=1}^{L}\|W_\ell\|_2\right)\|x - x'\|_2,$$

where matrix norms are spectral norms. Explain one learning-theoretic reason to control the product.

B. Deep Learning and Reinforcement Learning [33 points]

(a) **Batch normalization and dropout.** [9 points]

For one scalar feature in a minibatch $x_1, \dots, x_B$, let $q \in (0, 1]$ be a dropout keep probability.

(a) [3 points] Write the batch-normalization computation during training, including the learnable affine parameters and the update of running statistics.

(b) [2 points] State the inference-time computation and explain why using current test-batch statistics is undesirable.

(c) [2 points] For inverted dropout $\widetilde{h} = mh/q$ with $m \sim \text{Bernoulli}(q)$, compute its conditional mean and variance and state the inference rule.

(d) [2 points] Give one failure mode of batch normalization with very small batches and one suitable response.

(b) **A conditional variational autoencoder.** [15 points]

Let C be an observed class, and consider

$$p(c)p(z \mid c)p_\theta(x \mid z, c), \qquad q_\phi(z \mid x, c).$$

(a) [5 points] Derive a lower bound on $\log p_\theta(x \mid c)$. State the direction of every KL divergence.

(b) [4 points] Suppose

$$q_\phi = \mathcal{N}(\mu, \text{diag}(\sigma^2)), \qquad p(z \mid c) = \mathcal{N}(m_c, \text{diag}(\tau_c^2)).$$

Give the reparameterization and the closed-form KL divergence.

(c) [3 points] Describe one training step and the conditional generation procedure.

(d) [3 points] Define posterior collapse. Explain how an overly strong KL weight can cause it and give one diagnostic that uses both terms of the objective.

(c) **Three temporal-difference targets.** [9 points]

For a sampled transition (s, a, r, s'), a behavior policy chooses the next action from $\mu(\cdot \mid s')$.

(a) [3 points] Write the one-step targets for SARSA, Expected SARSA, and Q-learning. State what happens when s' is terminal.

(b) [3 points] Let $\gamma = 0.9$, $r = 1$, $Q(s', a_1) = 4$, $Q(s', a_2) = 2$, $\mu(a_1 \mid s') = 0.25$, and $\mu(a_2 \mid s') = 0.75$. If the realized next action is a_2, compute all three targets.

(c) [3 points] Which updates are on-policy and which can be off-policy? Explain the source of randomness in each target and why off-policy learning still requires coverage of useful state–action pairs.

C. Optimization Methods in Artificial Intelligence [33 points]

Let $C \subseteq \mathbb{R}^d$ be nonempty, closed, and convex, and let $F : C \to \mathbb{R}$ be convex. Assume that a minimizer $x^\star$ exists and that every subgradient used by the algorithm satisfies $g^k \in \partial F(x^k)$ and $\|g^k\| \le G$. Starting from $x^0 \in C$, projected subgradient descent with a constant stepsize $\eta > 0$ is

$$x^{k+1} = P_C(x^k - \eta g^k), \qquad k = 0, \dots, K-1,$$

where $K \ge 1$ is known in advance. Define $R := \|x^0 - x^\star\| > 0$ and $\bar{x}^K := K^{-1} \sum_{k=0}^{K-1} x^k$.

(1) [3 points] Give the definition of a subgradient and prove that $F(x) - F(x^\star) \le \langle g, x - x^\star \rangle$ for $g \in \partial F(x)$.

(2) [5 points] Using nonexpansiveness of Euclidean projection, prove the one-step inequality

$$\|x^{k+1} - x^\star\|^2 \le \|x^k - x^\star\|^2 - 2\eta\,[F(x^k) - F(x^\star)] + \eta^2 G^2.$$

(3) [7 points] Sum the one-step inequalities and derive a bound on

$$\frac{1}{K} \sum_{k=0}^{K-1} [F(x^k) - F(x^\star)].$$

Clearly identify the term that telescopes and the nonnegative term that may be discarded.

(4) [5 points] Use convexity to convert the result into a function-value bound for the averaged iterate $\bar{x}^K$. Explain why the same argument does not directly give the bound for the last iterate x^K.

(5) [4 points] Optimize the upper bound over the constant stepsize. Show that $\eta = R/(G\sqrt{K})$ gives

$$F(\bar{x}^K) - F(x^\star) \le \frac{RG}{\sqrt{K}},$$

and give a sufficient K for error at most ε.

(6) [5 points] Strong convexity alone does not make the constant-step subgradient method linearly convergent. Consider the one-dimensional strongly convex function

$$\Phi(x) = |x| + \frac{\mu}{2} x^2, \qquad \mu > 0,$$

with no projection and the natural subgradient away from zero. Assuming $0 < \eta\mu < 2$, construct a nonzero two-cycle of this iteration.

(7) [4 points] Compare the $O(1/\sqrt{K})$ guarantee above with the $O(1/K)$ rate of gradient descent on a convex smooth function. State exactly which missing assumption prevents the smooth proof from being applied here, and briefly interpret the importance of averaging.

D. Natural Language Processing [33 points]

1. Skip-gram with negative sampling. [7 points] For a center word w, a positive context word c, and negatives $n_1, \dots, n_m$, consider

$$\ell = -\log \sigma(\mathbf{u}_w^\top \mathbf{v}_c) - \sum_{i=1}^{m} \log \sigma(-\mathbf{u}_w^\top \mathbf{v}_{n_i}), \qquad \sigma(a) = (1 + e^{-a})^{-1}.$$

4 points Compute $\nabla_{\mathbf{u}_w} \ell$ and interpret every term.

3 points Explain why drawing negatives only from the empirical unigram distribution can overrepresent frequent words, and give one principled modification. Also explain the harm caused by a false negative.

2. Causal attention and autoregressive decoding. [7 points] Let $X \in \mathbb{R}^{L \times d}$ and $Q = XW_Q$, $K = XW_K$, $V = XW_V$. Define a causal attention layer precisely, including the softmax axis and mask. Then compare, for generation of T new tokens, (i) recomputing all keys and values at every step and (ii) using a KV cache. State the dominant time and cache-memory costs in terms of T and d (you may suppress the fixed prompt length and constant numbers of layers/heads). Explain what a KV cache does and does not change asymptotically.

3. Compute-constrained scaling. [8 points] Suppose the excess validation loss of a language model is

$$\mathcal{L}(N, D) - \mathcal{L}_\infty = AN^{-\alpha} + BD^{-\beta}, \qquad A, B, \alpha, \beta > 0,$$

where N is model size and D is the number of training tokens. Training compute is $C = \kappa ND$.

5 points Treating N, D as positive real numbers, derive the minimizing $N^*(C)$ and $D^*(C)$, including their powers of C.

3 points State the balance relation between the two loss terms at the optimum and give two reasons why extrapolating this law far beyond observed scales can fail.

4. Evidence-preserving long-document summarization. [11 points] A legal archive contains documents of up to 10^6 tokens. A user asks for a summary focused on a specified issue; every summary claim must link to supporting spans. Only $B \ll 10^6$ document tokens can be passed to the generator.

4 points Design a hierarchical selection and generation architecture. Define the intermediate representations and the token-budget constraint.

3 points Give a trainable objective for the selector when only document-level summaries are available, and explain how you would avoid a degenerate selector.

2 points Analyze the dominant time/memory costs of your design in terms of the number of chunks M, retained chunks k, embedding dimension d, and B.

2 points Specify evaluation metrics for coverage and faithfulness, and one adversarial failure test.